

A DISCUSSION OF INDIRECT SOUNDING METHODS

WITH SPECIAL REFERENCE TO THE DEDUCTION OF VERTICAL OZONE DISTRIBUTION FROM LIGHT SCATTERING MEASUREMENTS

S. TWOMEY AND H. B. HOWELL

Physical Science Laboratory, U.S. Weather Bureau, Washington, D.C.

[Manuscript received June 18, 1963; revised September 25, 1963]

ABSTRACT

Instability limits the usefulness of indirect sounding, i.e. the deduction of a physical distribution from a set of observations which represent an integral transform of the former. A method is presented which allows a stable, but smoothed, solution to be obtained in certain cases. As an illustration of the application of the method, the deduction of vertical ozone distribution from measurements of the spectral distribution of scattered ultraviolet radiation is discussed. Graphs showing results from several possible methods of inversion are included to show the difficulties associated with such indirect measurements.

1. INTRODUCTION

Primarily at the behest of Dr. Harry Wexler, one of the writers embarked a couple of years ago on an investigation which, it was hoped, would produce a new and superior method for deducing the vertical distribution of ozone from purely ground-based passive measurements.

The results up to now have shown that the projected methods, while comparable and potentially superior to existing ground-based passive methods, can never produce detailed distributions with the resolution of fine scale structure of which direct (balloon-borne, for example) sounding is capable. However the investigation has helped to give some insight into potentialities and limitations of indirect sounding methods, and has shown that there can be in the inversion process a fundamental instability which limits the resolving power of indirect soundings generally.

The phrase "indirect sounding" is used here to denote the use of measurements at a single place of a function $g(y)$ of a variable parameter y to infer the spatial distribution $f(x)$ of the desired quantity, there being no one-to-one correspondence between y and x . It may be mentioned that many methods which may be direct when considered in an idealized formulation can degenerate into the category of indirect sounding as a result of instrumental fallibility. The distinction between "direct" and "indirect" is, therefore, not always rigid.

2. THE GENERAL PROBLEM

In many instances a physical distribution $f(x)$ (for example, pressure versus ozone) can be shown to be related to another physical distribution $g(y)$ (for example,

spectral energy versus absorption coefficient) by an integral equation such as

$$g(y) = \int K(x, y) f(x) dx \quad (1)$$

From a mathematical point of view, an equation of this kind allows $f(x)$ to be described uniquely by $g(y)$, provided only that no two distinct functions of x lead to the same $g(y)$. In many instances—the Fourier transform being the most outstanding—either $g(y)$ or $f(x)$ may be most suited to the problem at hand, and one may shuttle back and forth between the function and its transform, for either $f(x)$ or $g(y)$ describe the function equally well.

Formally, therefore, one need only find a relationship of this kind to deduce spatial distributions from a set of observations at one place and one time. *Practically*, however, uniqueness of the transform is not sufficient, if one hopes to infer $f(x)$ from *measurements* of $g(y)$. The reasons for this are:

(i) Measurements can, at best, give the value of $g(y)$ at a finite number of values of the argument; these values are usually real, and often positive, values only. Thus $g(y)$ cannot be said to be defined in an analytical sense, and analytical inversions cannot be utilized directly.

(ii) Since there are always some uncertainties associated with measurements (however precise), $g(y)$ will not be known exactly at any point. In a graphical sense, one should draw a small circle about the measured point in the gy plane and assert only that the true $g(y)$ of equation (1) passes through each such small circle.

It thus becomes apparent that the crucial question is: Given a strip of finite extent ($c \leq y \leq d$) and finite width

in the plane of g and y , is the transform expressed by (1) such that all functions of x possessing transforms which lie within this strip in the interval $c \leq y \leq d$ themselves lie within a strip of finite width in the fx plane?

If the answer is in the negative the inversion of the transform is unstable, and a unique solution, or a solution with a predictable uncertainty, cannot be inferred from measurements of $g(y)$ alone.

Unfortunately most of the transforms encountered in practice are of this uncooperative character. This is usually associated with fixed, finite limits of integration and a kernel $K(x, y)$ which is smooth (often monotonic) with respect to x .

The case of the integral equation relating the spectral distribution of scattered light to the vertical distribution of atmospheric ozone furnishes a good example of the instability under discussion. It is not difficult to show that measurements of the scattered energy upward to a satellite instrument or downward to the ground are sufficient to derive values of the function $R(k)$, where

$$R(k) = \int_0^X e^{-kx} \frac{dp(x)}{dx} dx \quad (2)$$

If the solar zenith distance is Z and the instrument line of sight is inclined at an angle U to the vertical, then k is, for the satellite instrument $\kappa (\sec Z + \sec U)$ and, for the ground-based instrument $\kappa (\sec Z - \sec U)$, κ being in both cases the absorption coefficient; X is the total ozone in a column extending through the entire atmosphere. $p(x)$ describes the ozone distribution by giving the pressure level above which lie x units of ozone. k can be varied either by allowing the solar altitude to vary (Götz' Umkehr method), or by making observations at various wavelengths and thereby altering κ , which is a function of wavelength (fig. 1), or by varying the angle of observation, U . Thus in the case of "indirect sounding" of ozone from scattered light measurements, the kernel takes the form e^{-kx} and the limits of integration become 0 and X , respectively; the function sought is the integrand function $p(x)$ or $dp(x)/dx$.

The stability criterion under these conditions is most readily examined if the variables are changed to $\xi = \pi x/X$ and $\eta = kX/\pi$; considering dp/dx as an unknown function $f(\xi)$ of ξ , one may write

$$g(\eta) = \int_0^\pi e^{-\eta\xi} f(\xi) d\xi$$

and examine the stability of this transform. The function $f(\xi) - \frac{\xi}{\pi} [f(\pi) - f(0)] - f(0)$ vanishes at $\xi = 0$ and π and can therefore be written as a Fourier sine series (with coefficients b_m , say). By applying the integral transform to each term, there results:

$$g(\eta) = (1 - e^{-\pi\eta}) \left[\frac{b_1}{\eta^2 + 1} + \frac{2b_2}{\eta^2 + 4} + \dots + \frac{mb_m}{\eta^2 + m^2} + \dots \right] + (\text{a term independent of the } b\text{'s})$$

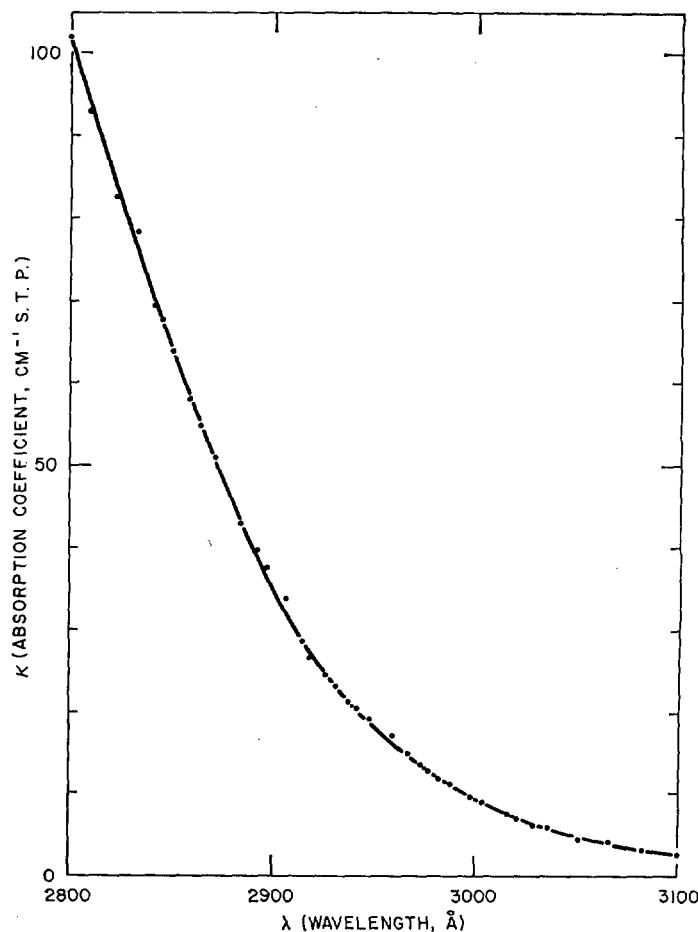


FIGURE 1.—Absorption coefficient κ as a function of wavelength λ .

Thus the transform is less and less sensitive the higher the frequency of the Fourier component. Hence two functions may differ greatly from one another and still possess transforms which differ very little. More specifically, the transforms of two functions f' and f'' differ absolutely by less than δ provided only that

$$\sum_m |(b'_m - b''_m)|/m < \delta$$

This inequality places no bounds on the norm $\sum_m (b'_m - b''_m)^2$.

Thus the transform is unstable; since it greatly diminishes the contribution of the higher order Fourier components, one might expect that the instability of the inversion might be manifested by the appearance of high frequency oscillations in the "solution" when $g(y)$ is to any extent imprecise.

In practice one cannot "solve" equation (1) in an analytic sense when $g(y)$ is measured at N points. The solution at best is a finite array of numbers (which may be values of $f(x)$ at selected points x_1, x_2, \dots, x_m , or coefficients in the expansion of $f(x)$ in terms of selected approximating functions). However, the instability just discussed still is present—and it is likely to be most

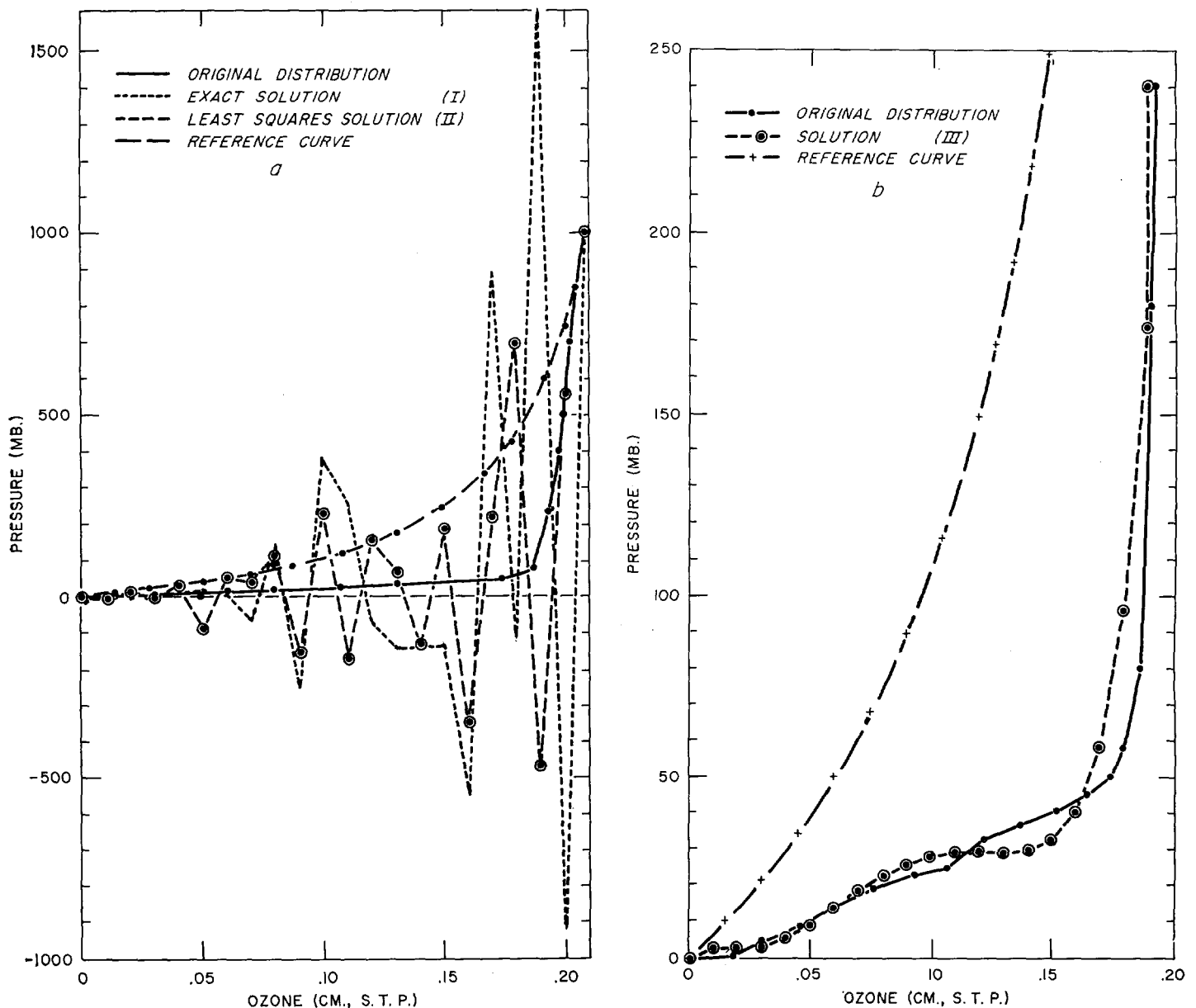


FIGURE 2.—(a) Comparison of exact solution (curve I) and least squares solution (curve II) with original hypothetical distribution of pressure versus total ozone above specified pressure level. (b) Comparison of solution by selection (curve III) with original hypothesized ozone distribution.

troublesome the greater the number of tabular points in x , since higher frequencies are thereby admitted. It is important to note that the instability of the inversion is a property of the kernel, not a result of the method of inversion.

In figure 2a a hypothetical ozone distribution is shown in terms of pressure versus total ozone above the specified pressure level. Obviously negative values of $dp(x)/dx$ are physically meaningless. For 28 values of k up to 30 (cm. STP) $^{-1}$, the quantity

$$R(k) = \int_0^x e^{-kx} \frac{dp(x)}{dx} dx$$

was computed to obtain the curve shown in figure 3. A quadrature formula of high accuracy was then constructed and a matrix \mathbf{A} thereby obtained such that

$$\mathbf{r} \approx \mathbf{A}\mathbf{f} \quad (r_i = R(k_i); f_j = p(x_j))$$

This (22-point) quadrature was accurate to better than $\frac{1}{4}$ percent for smooth, monotonic integrands.

When, however, the "solution" $\mathbf{A}^{-1}\mathbf{r}$ was computed, the result was ludicrous (curve I, figure 2a); when a least squares solution (curve II) was obtained, the result was no better. Yet when these distributions were inserted back into the integral equation they yielded for $R(k)$

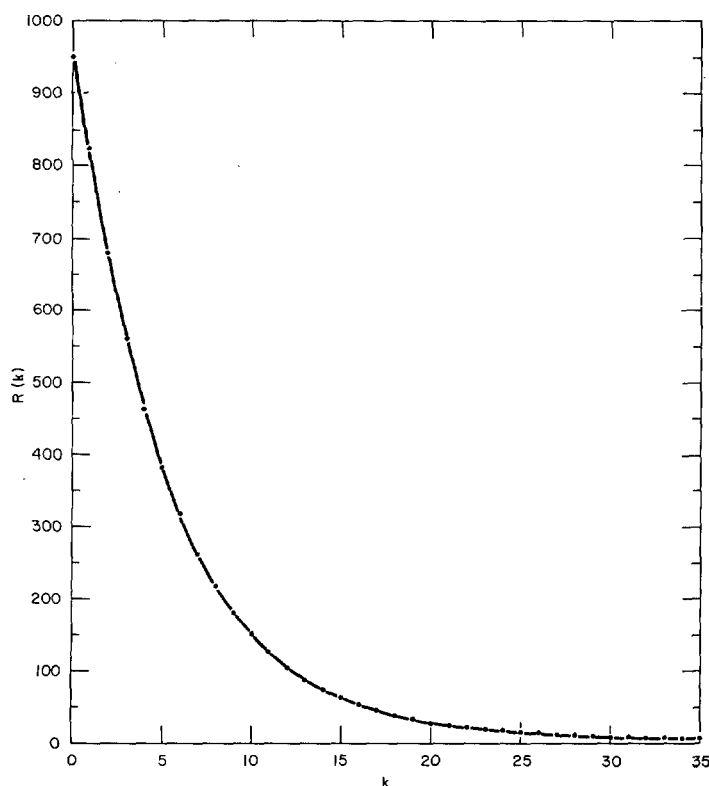


FIGURE 3.—The function $R(k)$ from which the solutions in figure 2 were obtained by inversion.

values almost identical with those which served as a starting point. Table 1 shows the original values $R(k)$, the values obtained by substituting the "exact" solution (I) into the integral, $R_I(k)$, and the corresponding quantities from the least squares solution, $R_{II}(k)$.

These results are a very vivid example of the instability of many integral equations. It is noteworthy that the degree of agreement found when a solution is inserted into the integral and the resulting transform compared with the given values of the transform *cannot* be used to assess the closeness of the approximation to that function which transforms exactly to those values. This merely restates what has been already demonstrated, but perhaps cannot be repeated too often, since many methods of inversion assume that nearness of the transforms in the gy plane implies nearness of the corresponding functions in the fx plane.

It may be remarked that in the matrix formulation $\mathbf{A}\mathbf{f}=\mathbf{g}$ corresponding to the integral equation, instability manifests itself in the guise of extreme skewness (non-orthogonality) of the quadrature matrix \mathbf{A} . \mathbf{A} (and, incidentally, the least-squares matrix $\mathbf{A}^*\mathbf{A}$ where \mathbf{A}^* is the transpose of \mathbf{A}) possesses several very small eigenvalues (10^{-6} and less); thus \mathbf{A}^{-1} and $(\mathbf{A}^*\mathbf{A})^{-1}$ possess very large eigenvalues. Hence the "exact" solution becomes, when an error ϵ exists in \mathbf{g} , $\mathbf{f}+\mathbf{A}^{-1}\epsilon$; the least squares solution becomes $\mathbf{f}+(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\epsilon$. But the large

eigenvalues of \mathbf{A}^{-1} and $(\mathbf{A}^*\mathbf{A})^{-1}$ can magnify ϵ to the extent that the error term dominates the inversion.

3. SOLUTION-BY SELECTION

The instability which has been demonstrated to exist in the case of the kernel e^{-xy} is so severe that it amounts to non-uniqueness for all practical purposes. Similar instability can be shown to exist in the case of other kernels which arise in similar indirect sounding problems.

The only practical, useful method of solution is to recognize that the equation

$$g(y) = \int K(x, y)f(x)dx + \epsilon(y)$$

has an infinity of possible solutions when $|\epsilon(y)|$ is everywhere $\leq \epsilon$, unless ϵ is exactly zero. It is possible uniquely to select from this infinite manifold of functions that function for which a specified measure of smoothness attains a maximum. This function can be regarded as the most probable solution, if the measure of smoothness has some *a priori* physical basis, for example.

In the case of the finite-difference equation analogous to the above integral equation, i.e.

$$\mathbf{A}\mathbf{f}=\mathbf{g}+\epsilon$$

it can be shown (Twomey [3]) that such constraints may be applied and a practically unique vector \mathbf{f}_1 computed by the equation:

$$\mathbf{f}_1 = (\mathbf{A}^*\mathbf{A} + \gamma\mathbf{H})^{-1}(\mathbf{A}^*\mathbf{g} + \gamma\mathbf{h})$$

where \mathbf{H} is a symmetric matrix, the form of which depends on the exact criterion or constraint applied; \mathbf{h} is a vector, which also depends on the constraint; and γ a Lagrangian multiplier, is determined by $|\epsilon|$. If, for example, the constraint requires that the selected "solution" be that which departs least from a specified curve, then \mathbf{H} becomes the unit matrix \mathbf{I} and \mathbf{h} is the specified curve in vector form.

In the case of atmospheric ozone distribution the general shape of the curve is predictable—particularly when the total ozone has been measured, for then the end point (P, X) of the $p(x)$ curve is also known. In figure 4 are plotted distributions from Dütsch [2]. It is obvious that once the end point is known, a curve can be drawn which will not lie too far from the true curve. For this reason the constraint just described was used to obtain inversions from the $R(k)$ data of figure 3. The inversion for $\epsilon=0.1$ percent is shown in figure 2b (curve III); the result of inserting this solution back into the integral equation is shown in table 1.

The reference curve which was used to get a constraint vector \mathbf{h} is shown in the figure. A curve closer to a typical ozone distribution could have been chosen, but this curve was used to show that the choice of the reference curve is not critical. It can also be shown, by using different

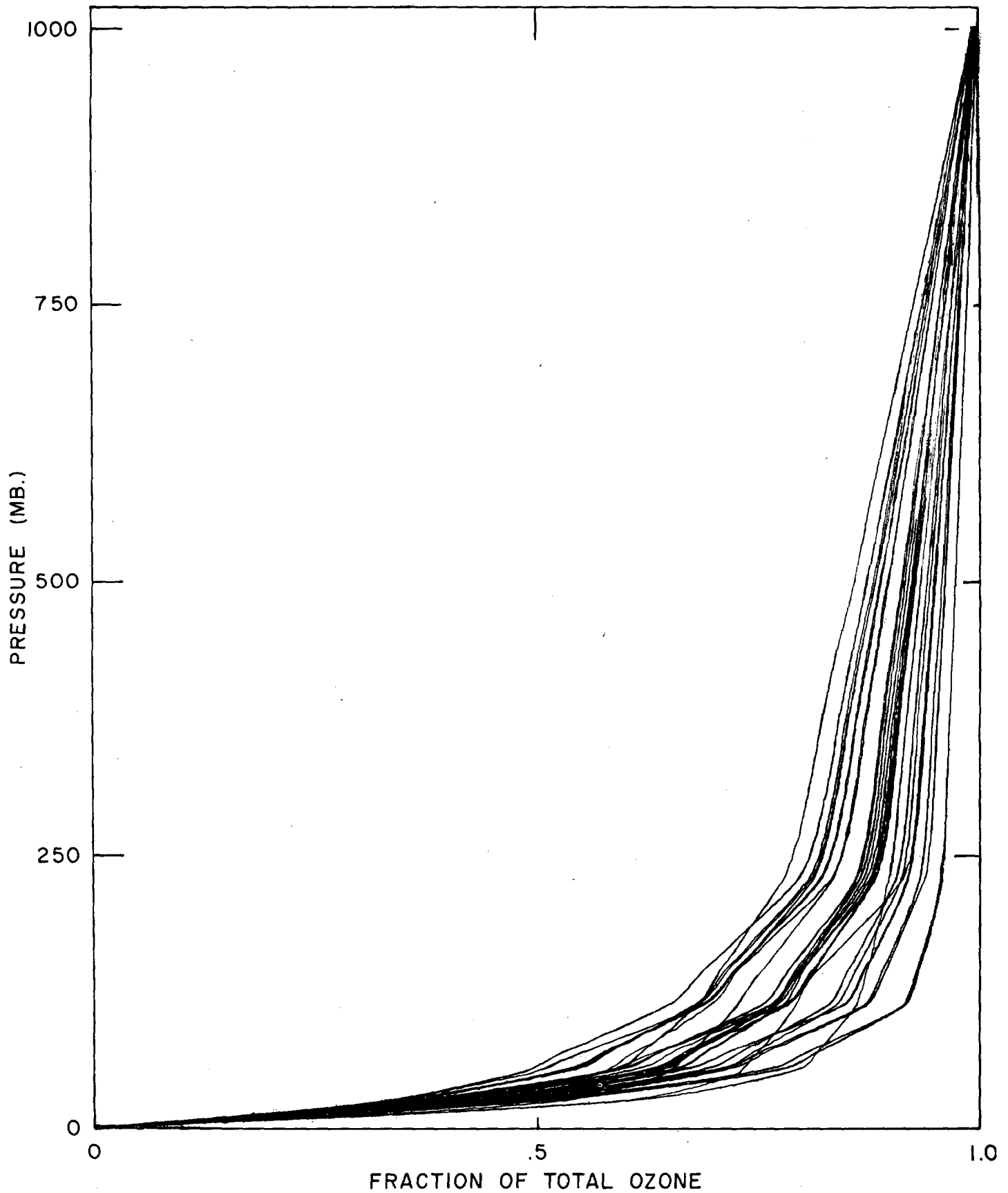


FIGURE 4.—Ozone distributions (from Dütsch [2]).

TABLE 1.—Comparison of original values $R(k)$ with values $R_I(k)$, $R_{II}(k)$, and $R_{III}(k)$, obtained by using the three solutions shown in figure 2.

k	$R(k)$	$R_I(k)$	$R_{II}(k)$	$R_{III}(k)$
11	125.7		125.7	125.7
12	105.2		105.2	105.2
13	88.29		88.29	88.31
14	74.33		74.33	74.35
15	62.80		62.80	62.82
16	53.26	53.26	53.26	53.27
17	45.36	45.36	45.36	45.36
18	38.80	38.80	38.80	38.80
19	33.35	33.35	33.35	33.35
20	28.82		28.82	28.81
21	25.04		25.04	25.03
22	21.88		21.88	21.87
23	19.24	19.24	19.24	19.22
24	17.01		17.01	17.00
25	15.14	15.14	15.14	15.13
26	13.56	13.56	13.56	13.55
27	12.22	12.22	12.22	12.21
28	11.08	11.08	11.08	11.07
29	10.11	10.11	10.11	10.10
30	9.276	9.275	9.275	9.271
31	8.556	8.556	8.556	8.555
32	7.934	7.933	7.933	7.937
33	7.392	7.392	7.392	7.400

values for γ that the exact choice of γ or ε is not at all critical, provided the value chosen for ε is not less than the true (and unknown) value of the norm of the error vector ϵ .

Not alone does this method of solution eliminate the spurious high-frequency oscillations, but it greatly simplifies the problem of correcting for attenuation by scattering, for example, since approximations can be introduced which allow for this effect at the expense of some increase in the inaccuracy of the quadrature formula. This increase can be "absorbed" by increasing γ , i.e. by allowing $|\epsilon|$ to take on larger values.

Interestingly enough the application of this kind of constraint in the solution process introduces a considerable numerical filtering, which discriminates against random errors to a marked degree. This arises from the fact that the manifold of functions of y which are transforms of smooth functions (or indeed, any continuous functions) of x in the interval $0 \leq x \leq X$ is a very restrictive manifold and a random error function $\delta(y)$ will be a sum of functions of which only a part will be within this manifold. There is the associated disadvantage that short-range fluctuations, even if present in the true $p(x)$, will be smoothed out in the solution. However, since the measured quantity $R(k)$ is so insensitive to such fluctuations, their retention in a solution is meaningless; but it must be emphasized that the solution obtained by this process is a smoothed $p(x)$ and cannot be used as evidence for or against any fine structure in the distribution.

4. DISCUSSION

The preceding analysis and numerical results exemplify the difficulties which are encountered in indirect soundings. That these difficulties are fundamental and not procedural is apparent; that similar difficulties will be encountered whenever the kernel is smooth can be demonstrated, for example by elaborating the Fourier method of section 2.

Since optical transmission functions tend to be smooth, it would seem that all methods of indirect sounding by optical means will be similarly troubled and will be specifically incapable of resolving fine structure in the sounding. That little significance can be attached to fine structure in ozone distributions computed from Umkehr data—in particular narrow sharp peaks in ozone concentration, which affect very slightly the integrated ozone in a column—goes without saying.

It is worth noticing that the "resolving power", in the sense of the number of independent points or parameters which can be deduced, is determined by the number of eigenvalues of the matrix $\mathbf{A}^* \mathbf{A}$ exceeding a definable lower limit, which depends on the accuracy of measurement, of the quadrature, and of any other approximations made. But the eigenvalues of $\mathbf{A}^* \mathbf{A}$ are merely the extremal values of the absolute magnitude of a normalized arbitrary linear combination of the row vectors of \mathbf{A} ; i.e. they are measures of independence (Courant-Hilbert [1]) of the kernel functions. Increasing the number of values of the parameter y within a fixed interval does not, beyond a certain point, add to the effective number of independent observations; on the other hand, if points can be added outside the previous interval, the degree of independence increases and the "resolving power" benefits. In the case of the ozone distribution problem, for example, this would involve measuring an increasingly smaller amount of energy. Thus the effective "resolving power" is dictated by a combination of factors, some purely instrumental.

REFERENCES

1. R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Interscience Publishers, Inc., New York, 1953, 1st ed., vol. I, Chap. 2, pp. 61–62.
2. H. Dütsch, "Vertical Ozone Distributions from Umkehr Observations," *Arkiv für Meteorologie, Geophysik und Bioklimatologie*, Serie A, vol. 11, No. 2, 1959, pp. 240–251.
3. S. Twomey, "On the Numerical Solution of Fredholm Integral Equations of the First Kind by the Inversion of the Linear System Produced by Quadrature," *Journal of the Association for Computing Machinery*, vol. 10, No. 1, Jan. 1963, pp. 97–101.